Imperial Business Analytics

Modelling Conditional Probability of (Re-) Emerging Infectious Diseases

Data Spark Consulting Project for ProMED

Dr. David Birch
Valeria Cortez Vaca Diez
Valentin Courgeau
Haochen Song
Kangyan Zhao
Tobey Agbo-ola

Imperial College London
April 2018

Imperial Business Analytics

Table of contents

Imperial Business Analytics

1. Executive Summary

Preventing the spread of infectious diseases is a major concern for the public health and the global economy. If we could forecast where a communicable disease would spread next, we could then intervene earlier to prevent the loss of life and economical damage.

ProMED is is an Internet-based reporting system dedicated to rapid global dissemination of information on outbreaks of infectious diseases and acute exposures to toxins that affect human health and others. This organisation provided us a list of over 36,000 outbreaks alerts of infectious diseases ranging from 2002 to 2017 around the globe.

The goal for this project was to calculate the probability of an infectious disease outbreak in country B, given that country A is infected by the same disease. This would allow ProMED and its network to:

- identify countries that are more likely to infect others
- set early actions to prevent the spread of diseases in countries that could get infected

We focused on Yellow Fever as there have been recent outbreaks and these results can have an immediate impact. We also looked into African Swine Fever to understand if this model could be extended to infectious disease affecting animals.

We created disease-specific datasets that included for each day, every country that had presented an alert. For each of these observations, we looked for the presence of alerts for all other countries. This process created country pairs (A, B) and a binary variable to identify if there had been an alert in country B, given that there was an alert in country A.

For each of these pairs, we added information on country factors, time difference between reports, environmental factors and vehicles of travel. We used these as regressors to model the probability of an outbreak in country A, given that country B is infected by the same disease.

Findings:
After trying different models, we found that probit regression model was the most appropriate model used to calculate the probability and simulated disease transmission. From the model we could see that some countries were more likely to infect others while other countries were more vulnerable to be infected by others.

For yellow fever, the countries that were more likely to infect others were usually wealthy and well connected by air traffic. Countries that were more vulnerable to contract an infectious disease usually included developing nations with lower levels of public health and socio-economic and political stability. Nevertheless, we could also see that Schengen countries were more vulnerable compared to other wealthy countries, suggesting that open borders may have a strong impact on the spread of communicable diseases.

Recommendations

We recommend ProMED to use these findings to identify most infectious countries and most vulnerable countries to be infected. In this way, they can inform their partners on how to best focus their resources on such countries in order to prevent the spread of these diseases.

Using the probit model, ProMED can also make prediction on possible changes as the improvement of public health or political stability. This can provide data evidence to take action in certain areas that can be developed.

2. Client Introduction

As a non-profit organization founded in 1999, ProMED was developed to monitor emerging diseases based on Internet reporting system.

The main targets include infectious diseases and acute exposures to toxins that affect human health, plants and animals. ProMED receives and distributes all different kinds of information sources, including media reports, official reports, online summaries, local observers and others from organizations or individuals (ProMED, 2010).

Currently, ProMED has 60,000 members worldwide and reaches in more than 185 countries. ProMED will collect these disease outbreak information from report posted and articles published within 24 hours. There are some specialist who will double check whether the alert is of a (re)emerging disease or not. If it is not, they will decide not to publish this alert. These published alerts will be distributed to doctors that recommend people about diseases when they travel, 90k subscribers, professors, authorities (e.g. minister for health) and organizations (e.g. WHO).

By providing disease outbreak information, ProMED can give early warnings to relevant countries, so that they can take timely initiatives to prevent epidemic transmission and save more lives.

3. Dataset

3.1 Description

The ProMed dataset contained 10 columns in total that are described in the table below:

alert_id	the unique ID for the ProMED report
alert_meta_id	the unique ID for the meta tag
issue_date	the date the post was made on ProMED in the Eastern US time zone
place_name	the name of the place where the outbreak was located
name	the country name associated with the place_name
disease_name	the name of the disease associated with the meta tag
species_name	the species affected by the disease
href	a stem which can be used to retrieve the original post from the ProMED-mail website
summary	the subject line of the email sent for this ProMED post
content	the complete text of the email sent for this ProMED post

Table 1: Attributes of ProMed's dataset.

Country	Brazil
City	São Paulo
issue_date	10 April 2018
species_name	Human
disease_name	Yellow Fever
content	A white spot shrimp disease outbreak in Brazil has led ...

Table 22: Subset of ProMed's dataset.

3.2 Exploration

The subset we used contained 36,133 alerts within the year of 2002-2017 and covers 208 countries in total. Details of both time trend exploration and

cross-sectional exploration will be shown in the following.

a) Time Trend

Figure 1 shows the total amount of alerts for 2002 to 2017. There is a gradual increase from 2518 to 5018 until year 2009 and then sharply drops to 967 at year 2013. After year 2014, the amount of alerts remained at around 4,500.

Figure 1: *Time trend - number of alerts*

Figure 2 shows the number of countries included during the time period. Similar to the trend of alerts, there is a decrease of value from year 2009 to 2013. After reaching to the bottom of 114 in 2013, the number of country increased to 190 and then remained constant at around 170.

Figure 2: *Time trend - Number of countries*

To sum up, during year 2009 to 2013, both the total number of alerts and countries included in the dataset decrease sharply. Either lack of data or lack of actual reports during this time could be a reason.

b) Cross-sectional Exploration

Figure 3 below indicates number of times that each country is mentioned among all the alerts. From the graph we can see United States has the highest number of alerts followed by China and India.

In order to further explore the geographical distribution of these alerts, we created a distribution map in Figure 4. The size of red point in Figure 4 indicates the number of total alerts in each country. We can see from the distribution that ProMED's alerts almost cover all areas around the world and similar to Figure 3, United States, China and India are still the leading countries, followed by United Kingdom and Canada.

Figure 3: *Number of alerts by country*

Figure 4: Number of alerts by country - *Geographical map*

Figure 5 below shows the rank of number of disease mentioned. Avian Influenza has the highest number of total alerts but the gradient color also indicates the number is diseasing as time passes. Contrary to Avian Influenza, Dengue (rank 2nd) experiences an increase trend as time goes on. More Dengue alerts emerge in these

years. Some diseases such as Influenza H1N1 and SARS only occured in the previous years, which indicates the probability of reemerging is low.

Figure 5: Number of alerts by disease

In conclusion, the above five graphs show general information of the dataset we have from Promed in different dimensions. This dataset is going to be used in our project.

4. Project Description

4.1 Goal

As a wicked problem, monitoring the emergence and spread of infectious disease has been at the centre of the world. The recent outbreak of Ebola attracted international authority's attention about infectious disease monitoring and highlighted its significant value. The outbreak of Ebola virus in west Africa was the largest outbreak of infectious disease, a total of 28,616 suspected cases and 11,310 deaths were reported according to WHO (WHO, 2016). Active monitoring can help prevent and contain outbreaks of rapidly spreading emerging pathogens that pose threat to public health. For example, the U.S. Centers for Disease Control and Prevention (CDC) regulated that individuals under active monitoring were asked to contact local health authorities about their health condition every 21 days (Reich et.al, 2017).

Due to the limited resources, ProMED cannot allocate a specialized data analysts team to gain more insights from free-text reports, which significantly influence their data extraction ability. Using limited data, ProMED can only set early warnings for relevant countries to prevent the spread of diseases. Our goal is to help them take a step further to build a more quantitative alert platform.

Our project is focused to help ProMED provide recommendations on potential "next targets" of a particular disease given the news of an outbreak. We aim to achieve this by calculating the conditional probability of an infectious disease outbreak in country A, given that country B is infected by this same disease within a time period.

We collected country-factors that allow communicable diseases to spread more easily. These factors can be channels of connections (e.g. connecting flights) or factors of vulnerability (e.g. health care services).

By merging surveillance datasets and running statistical and machine learning models, we aim to help ProMED understand infection transmission in a globalised and well connected world. In this way, ProMED can identify countries that are more likely to infect others in advance and communicate with partners in an engaging way.

4.2 Limitations

4.2.1 Data unavailability of uncommon diseases

ProMED publishes data on emerging and re-emerging diseases. This implies that there are alerts on infectious diseases that are not published given that they are perceived common. The definition of 're-emerging' is subjective and is defined by ProMED network. We are unaware of the percentage of alerts that are not published.

That means that the dataset we were working with did not include an unknown group of infectious diseases that are considered common.

4.2.2 ProMED network's evolution

Our dataset of alerts spans from 2002 until 2017 and in the meantime, ProMED's network has evolved and widen. In our approach, we consider the network stationary, i.e. has remain the same through time, and perfect in the sense that every infectious outbreak has been reported. This means that we do not consider tackling the famous *under-reporting problem* in disease outbreak research. Indeed, any quantitative approach requires a benchmark to which we could refer to and the only one available is the World Health Organization (WHO) which only proposes yearly statistics on a restricted number of diseases. This would mean that we could not harness the small time granularity of our dataset and is the reason why we decided not to take this angle.

4.2.3 Number of alerts by country

The main contributor to ProMED alert system are the United States of America. Indeed, approximately 25% of them come from different organisations based in the US about outbreaks on American soil. We need to consider the impact of US carefully when looking into the results as it will inevitably induce bias in our estimations. We explain this by the abundance of the health organisations as well as research centres all around the country.

At a much smaller scale, we also see a large number of alerts coming from China, India, UK, Australia and Canada. We also need to pay attention the latter three, given that they present smaller populations than the first two.

5. Literature Review

Preventing the spread of infectious diseases is a major concern for the public health and the global economy.

The forecasting of diseases is particularly challenging due to diverse contact patterns (Mossong et al. 2008), nonlinear transmission dynamics (Woolhouse 2011), seasonal variations (Metcalf et al. 2009), spatial dynamics through human travel (Held et al. 2017) and data availability (Eisenstein 2018).

There have been attempts to use Big Data solutions to model the spread of diseases with online data from searches and news media; however, these have suffered from seasonality errors from “media hypes” (Eisenstein 2018) - as an example, we can cite the ‘Bieber fever’ about the singer Justin Bieber as one of those pitfalls.

There is a need to know where diseases can spread, so that countries and international organisations plan accordingly early detection, hospital resources and prevention campaigns (Shaman 2016).

Held et al. 2017 suggest that infectious disease forecasting should be probabilistic in nature or take into account temporal dependencies inherent to communicable diseases, spatial dynamics through human travel and social contact patterns between age groups.

As part of the temporal dependencies, the RAND Corporation created an Infectious Disease Vulnerability Index to help identify and raise awareness of those countries that might be most vulnerable to infectious disease outbreaks (Moore et al. 2017).

The dependencies included the following seven country-specific domain factors:

- Demographic Domain
- Health Care Domain
- Public Healthcare Domain
- Disease Dynamics Domain
- Political Domestic Domain
- Political International Domain
- Economic Domain

Each domain contributes to a higher lower or lower score that impacts the vulnerability of a country contracting an infectious disease.

Moore et al. found that the seven of the 10 most vulnerable countries were in conflict-zones. Moreover, 24 from of the 30 most vulnerable are on a continuous belt from the edge of West Africa to the Horn of Africa in Somalia.

The least vulnerable countries usually present robust democracies and stable economies and health systems.

6. Methodology

6.1 Dataset building

ProMED provided us a list of over 36,000 outbreaks alerts of infectious diseases ranging from 2002 to 2017 around the globe.

We created disease-specific datasets that included for each day, every country that had presented an alert. For each of these observations, we looked for the presence of alerts for all other countries. This process created country pairs (A, B) and a binary variable to identify if there had been an alert in country B, given that there was an alert in country A.

For each of these pairs, we added information on country factors, time difference between reports, environmental factors and vehicles of travel. We used these as regressors to model the probability of an outbreak in country A, given that country B is infected by the same disease. The factors are explained in the section 6.2.

We identified and accessed over 30 external datasets covering over 10 years of data that were used for the country factors, environmental factors and vehicles of travel. This also involved cleaning each dataset and filling missing values based on the average value of the corresponding sub-region. We also normalised the datasets and formatted as necessary.

6.2 Regressors

As mentioned earlier, we added country-specific regressors as well as time difference between alerts for each country pair (A,B).

Each regressor is explained in detail in the sections below.

6.2.1 Magnitude

In order to be able to analyse the magnitude of an outbreak from ProdMED's 'free text information' alerts, we employed Natural Language Processing (NLP) technique using SpaCy for implementation. SpaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python (Anon n.d.).

To extract number of cases (actual number of people affected or deaths) using the SpaCy library features, each alert, was tokenized first, segmenting the text into sentences and further into tokens (words) within that sentence. This was then followed by part-of-speech (POS) tagging to assign word types to each token (word).

Part-of-speech tags are the properties of the word that are defined by the usage of the word in the grammatically correct sentence like noun, verb, cardinal number, etc.

For every token (word), POS tagged '**Cardinal number**' e.g. '10' or 'Ten'. We then applied dependency, which assigned syntactic dependency labels, describing the relations between the cardinal number tokens and the subject or object token (word) to form a 'phrase chunk' for that number.

When generating the 'phrase chunk' around every number, a filter was applied to eliminate phrase chunks without words like 'cases', 'death' and 'people' etc and all their synonyms. Then, the number of cases were extracted from the remaining phrase chunks for each alert.

Unfortunately, after manually checking the extracted number of cases, we found out there were too many repetitions of the same number of cases or of previous cases.

In order to address these exaggerated numbers, we attempted to use "fasttext" deep learning tool for text classification by manually labelling some phrase chunks as sample data and applying the deep learning tool. Due to times constraints, we were not able to apply this and had to drop the magnitude regressor from the analysis.

6.2.2 Vehicles of travel

It seems very sensible to think of vehicles of travel for infectious diseases. Indeed, we all have this idea that diseases spread via different vectors such as animals, air currents, population movements, etc. But are these reliables indicators to model disease outbreaks? From our discussions with the client and advisors, it does play a important role and is usually considered for modelling purposes. For this project, we wanted to consider two main causes of travel: shared borders and air travel.

- *Shared borders:* our assumption was that terrestrial links are a good indicator of how close two countries are and how likely people are to cross go from one country to the next;
- *Air travel:* A fast way for an infected individual to spread the disease elsewhere is via air travel as it allows to reach far away countries within a day. Another risk in the contamination of travellers during the flight itself due the proximity of passengers (Kenyon et al. 1996);

6.2.3 Time difference

To include a temporal component to our study we first considered the series of alerts as a time series which a series of data points indexed by time. The issue is that this method will only be useful for disease for which there are many alerts (e.g. seasonal influenza) and cannot be extended further to uncommon/rare diseases. Our idea was to have a regressor about the time difference between the alert we have received and the alert horizon for which we would like to predict. In concreto, we use a

maximum threshold of 90 days this time difference. We consider that two countries cannot infect one another after this said threshold as new alerts will come in if the situation gets worse. Ideally, we would like to know the potential new targets within a month of a newly arrived alert.

6.2.4 Country factors

We used the 7 domain factors proposed by (Moore et al. 2017) as a basis for the regressors (See Figure 6).

Figure 6: Domain factors impacting the nation's ability to prevent or contain a disease outbreak as proposed by Moore et al. 2017

Our adapter version, conveys the following:

- **Demographic Domain:** Higher score indicates that an infectious disease is more likely to spread to other countries due densities and more human contact.
- **Health Care Domain:** Higher score means that the country can better contain an infectious disease with better health care services.
- **Public Health Domain:** Higher score means that the country is less likely to transmit the disease to more people due to better public health infrastructure.

- **Disease Dynamics Domain:** Higher score indicates that a country is more likely to transmit and spread a disease.
- **Political Domestic Domain:** Higher score means that a country is better to cope with an infectious disease due to a strong government and political stability.
- **Political International Domain:** Higher score indicates that a country has the international support to better cope with an infectious disease.
- **Economic Domain:** Higher score means that the country can better contain a disease due to economic strength.

Each domain is based on different datasets. Each dataset is scaled between 0 and 1 and it is modified when necessary to be in the right direction. The direction is simply whether a low value represents no impact on vulnerability and high value the opposite.

(Moore et al. 2017) created each domain using weights for each dataset score based on rigorous literature review and extensive experiences in epidemiology.

The tables 1 to 7 below describe the factors, hypothesis, measures and weights as in the study of (Moore et al. 2017). The greyed areas are for the datasets in "Measure" that we were not able to get or were not comprehensive enough. The direction column is the direction we decided to use.

Though for most datasets we had information available for most countries, we filled missing values by taking the year average of the country's subregion.

We used the weights as in the research done by (Moore et al. 2017).

Table 3: Demographic Domain

Factor	Hypothesis	Measure (Dataset)	Weight	Direction
Population density	A country with higher population density is more susceptible to the spread of emerging infectious diseases via overcrowding	Persons per square km	1	High= bad Low= good
Urbanization	A country with densely populated urban areas is more susceptible to the spread of infectious diseases via	Percentage of persons living in urban areas	1	High= bad Low=

	overcrowding and direct or indirect contact with numerous persons			good
Human population growth	A country with higher growth in population is more susceptible to the spread of emerging infectious diseases via overcrowding	Annual population growth rate (average annual percentage change in population)	1	High= bad Low= good
Education/ literacy	A country with high rates of literacy and education is less susceptible to the spread of emerging diseases via risky behaviors that may increase exposure	Adult literacy rate Adult female literacy rate		
Population mobility	A country with high migration and mobility of peoples is more susceptible to the spread of infectious diseases	Net migration rate (average annual number of migrants per 1,000 people)		

Table 4: Health Care

Factor	Hypothesis	Measure (Dataset)	Weight	Direction
Medical care expenditures	A country with greater spending on health is better able to limit infectious disease outbreaks	Percentage of GDP spent on health Health expenditure per capita	1 1	High= good Low = bad
Health status/ outcomes	A country with worse health status, with infant mortality rate as a proxy, reflects less ability to deliver services and in turn is less able to respond effectively to prevent or limit infectious disease outbreaks	Infant mortality rate (number of deaths in <12 months per 1,000 live births)	1	Data flipped, so that High= good Low = bad
Medical care workforce	A country with more health care providers is better able to limit infectious disease outbreaks	Number of physicians per 1,000 population Number of nurses and midwives per 1,000 population		
Medical care infrastructure	A country with a better medical infrastructure is better able to respond to limit infectious disease outbreaks	Hospital beds per 1,000 population Health posts per 100,000		

		population Health centers per 100,000 population Hospitals per 100,000 population		

Table 5: Public Health

Factor	Hypothesis	Measure (Dataset)	Weight	Direction
Health service delivery	A country that is better able to deliver basic primary care services is also better able to respond to limit the spread of infectious disease outbreaks	Percentage coverage with third dose of DTP vaccine		
		Percentage coverage with first dose of measles vaccine		High= good Low = bad
Water, sanitation, and hygiene infrastructure	A country with more widespread availability of potable water, sanitary conditions, and proper hygiene is better protected against the transmission of some infectious diseases	Population using improved drinking-water sources (%) Population using improved sanitation facilities (%)		High= good Low = bad
Basic public health infrastructure	A country with a strong public health infrastructure—e.g., having a national public health institute— is better able to prevent and respond effectively to limit infectious disease outbreaks	Country is member of the International Association of National Public Health Institutes		
Composite IHR core capacity score	A country with stronger IHR core capacities is better able to prevent and respond effectively to limit infectious disease outbreaks	Arithmetic average of score across all IHR scores		
GHSA action packages	A country that is committed to lead or contribute to a GHSA action package will be better able to contain infectious disease outbreaks	Country leading or contributing to >1 GHSA action package		

Table 6: Disease Dynamics

Factor	Hypothesis	Measure (Dataset)	Weight	Direction
Precipitation/ rainfall	A country with greater precipitation can have greater transmission of water- and vector-borne diseases because of the effects of precipitation on the replication and movement (and perhaps evolution) of disease microbes and vectors	Average rainfall per year (mm)	Used as single regressor	High= bad Low = good
Temperature	A country with higher temperatures can have greater transmission of water- and vector-borne diseases because of the effects of temperature on the replication and movement (and perhaps evolution) of disease microbes and vectors	Annual average temperature	Used as single regressor	High= bad Low = good
Changes in land use	Increasing anthropogenic activities is associated with increased susceptibility to and likelihood of emergence of zoonotic infectious diseases—either by increasing proximity or, often, by changing conditions that favor an increased population of the microbe or its natural host	Agricultural land (%) Forest area(%) Global deforestation rates (%)	1 1 1	High= bad Low = good

Table 7: Political-Domestic

Factor	Hypothesis	Measure (Dataset)	Weight	Direction
Governance	A country with a competent and strong government is better able to contend with an infectious disease outbreak	Worldwide Governance Indicators Government Effectiveness Index Worldwide Governance Indicators Regulatory Quality Index Worldwide Governance Indicators Rule of Law Index	1 0.5 0.75	High= good Low = bad
Corruption	A country with greater corruption has worse health outcomes and greater vulnerability to infectious disease outbreaks	Transparency International Corruption Perceptions Index	1	High= good Low =

				bad
Government stability	State fragility increases vulnerability to infectious disease outbreaks, while infectious disease outbreaks can exacerbate existing state weaknesses	Fund for Peace Fragile States Index	1	High= good Low = bad
Presence of conflict	Political stability—absence of conflict and state fragility—is associated with better ability to deliver health care and better health outcomes	Worldwide Governance Indicators Political Stability and Absence of Violence Index (World Bank)	0.75	High= good Low = bad
Service Provision	A country with greater stability and better quality of services has fewer barriers (geographical, financial, personnel, and access) to health care for marginalized populations	United Nations Development Programme Human Development Report Health Systems Survey		
Decentralisation	Various dynamics of decentralization (fiscal, political) are linked with positive health outcomes	World Bank decentralization index		
Democracy	A country with a more democratic and legitimate government is better able to contend with an infectious disease outbreak	Polity IV Project Democracy Index		
Human rights	A worse human rights record is linked with worse health performance	Amnesty International Political Terror Scale U.S. Department of State via Amnesty International Political Terror Scale		

Table 8: Political-International

Factor	Hypothesis	Measure (Dataset)	Weight	Direction
Aid support	States receiving more donor aid are better able to ensure health system functionality	World Bank Net Official Development Assistance per capita	0.75	High= good Low = bad
Aid dependence	Countries with a high proportion of donor aid are less able to deal with health	World Bank Net Official Development Assistance received (% gross national	1	Data flipped, so that

	emergencies on their own and therefore are more vulnerable to infectious disease outbreak	income)		High= good Low = bad
International organisation support for health	International organization and bilateral support to developing countries should lead to health sector strengthening and better resiliency against and response to infectious disease outbreaks	Development assistance for health per capita	0.75	High= good Low = bad
Aid continuity	Consistent, predictable funding support can promote better infectious disease control through stronger health systems	Lagged correlation between foreign aid and foreign direct investment		
International organisation support	Greater involvement, funding, and assistance by intergovernmental or bilateral partners will lead to more-effective detection and control of infectious disease outbreak	United Nations Development Programme recipient funding by country per capita		
Collaboration	Collaboration across governments, donors, and NGOs in program design and implementation is associated with better health systems and infectious disease control	Involvement with multilateral institutions (Jane's)		

Table 9: Economic

Factor	Hypothesis	Measure (Dataset)	Weight	Direction
Economic strength	A strong economy is associated with better health outcomes (lower infant mortality and longer life expectancy) in all countries	GDP per capita	0.75	High= good Low = bad
Economic growth	Greater economic growth has led to significant gains in control of infectious disease outbreaks even in countries with weak	GDP per capita growth rate	1	High= good Low = bad

	institutional environments; the gains from economic growth flow directly into health gains, up to a certain threshold of development			
Economic development	Countries with stronger economic development have greater access to diagnostic resources, making these countries more able to detect and respond to infectious disease outbreaks	United Nations Development Programme Human Development Index	0.75	High= good Low = bad
Partner-nation communications infrastructure	Good communications infrastructure makes it easier to deliver information about infectious disease and control measures to the population and outlying authorities	Cell phone subscriptions per 100 people Internet users per 100 people	0.5 0.25	High= good Low = bad
Partner-nation transportation infrastructure	Good transportation infrastructure makes it easier to deliver needed medical supplies to a country and to distribute them throughout the country	World Bank Logistics Performance Index Percentage of paved roads (of total)		
Technological sophistication	Greater technological penetration and sophistication are associated with better infectious disease control	Knowledge Economy Index		

6.3 Models

We used Probit Regression with Oversampling, Random Forest Classification and LogitBoost as models.

We splitted 50% of the data as training set, 25% as validation and 25% as test. We trained the model in the training set, tuned parameters and selected model using the validation dataset and compared results across different model families with the testing set.

At the end, we created a dataset for visualisation purposes. This dataset included every country as source country A, paired to each other country. For each pair, we

added their corresponding country metrics. We used 30 days as time difference and used January as the month to calculate the temperatures and rainfall of each country.

In terms of metrics, we got the best results from the Random Forest Classifier using the test data. However, it seemed that due to the nature of the dataset, this classifier was overfitting the data. When we visualised the results in an interactive map using the visualisation dataset, we noticed due to domain knowledge that likelihoods relations between countries were not reasonable. We explain this by the nature of the data: we work on disease-specific datasets, from which a small minority is about infected countries. Our models rely on very few data points to model infection mechanics thus extrapolating it with an unseen or fictitious origin country is difficult hence the poor performance. Random Forests' performance relies a lot on hyperparameters such as the tree length and set of regressors involved in addition to the dataset size. This has made the approach unlikely to scale well to a larger selection of disease or more uncommon ones.

Probit Regression with Oversampling provided more justifiable results. Moreover, this model outputs more quantitative explanatory information that ProMED could use for different purposes.

LogitBoost did not achieve as good results as Probit Regression with Oversampling and it was therefore dropped. Indeed, it was performing very well on guessing the main class of non-infection but struggled to guess infectious states as it performs worse than giving the answer perfectly at random.

The sections below explain in more detail the results of each model.

6.3.1 Probit Regression with Oversampling

Using the regressors as explained earlier in section 6.2, we modeled the likelihood that a disease could spread to a different country. We got the results as seen in Table 10.

With exception to Political Domestic Domain (Country B) and Rainfall (Country A) for Yellow Fever and Economic Domain (Country A) for African Swine Fever, all the regressors were significant.

Table 10: Probit regression for each disease

For the models of these two diseases, the *accuracy* was around from 0.716, *sensitivity* was 0.716 for both models and *specificity* varied from 0.683 to 0.723 (See Table 11)

Table 11: Confusion matrix and metrics for Probit Regression ('Positive' Class=0)

Yellow Fever

Pred / Actual	0	1
0	17,621	1,147
1	7,008	3,099

Accuracy:	0.717
Sensitivity:	0.716
Specificity:	0.723

African Swine Fever

Pred / Actual	0	1
0	7,935	63
1	3,145	136

Accuracy:	0.716
Sensitivity:	0.716
Specificity:	0.683

We will now look into the marginal effects of the probit regression of yellow fever.

We will start with the country indices from the (Moore et al. 2017) study as these range from 0 to 1.

We can see that the economic domain of the source country has the largest effect on the likelihood of a country getting infected. This may suggest that countries with better economies can better cope at retaining a communicable disease.

Other important factors of this type include include the demographics domain in country B, public health in country B and rainfall in country B.

We can now look into the regressors not proposed by (Moore et al. 2017). We see that if countries A and B are neighbours, the likelihood of infection increases by 0.311.

Every additional airport between country with connecting flights from A and B increases the likelihood by 0.375.

As time passes, the likelihood of infection is as expected smaller. Nevertheless, the effect is very small.

Table 12: Marginal effects for Probit Regression of Yellow Fever

We will now proceed with the marginal effects of the probit regression of the African Swine Fever.

From the country indices, we can see that the level of rainfall in country B has a strong effect on decreasing the likelihood of transmission. An increase of 0.1 in rainfall decreases the likelihood by -0.233.

Interestingly, a higher value in the demographics domain decreases the likelihood, with more emphasis if it is in country A. It is possible that as urbanisation and the human population increases, there is less agricultural activity.

The disease dynamics domain has also a strong impact on increasing the likelihood. This may be expected as this domain takes into account the changes in land use that directly impacts the likelihood of emergence of zoonotic infectious diseases.

We also see that if countries A and B are neighbours, the likelihood of infection increases by 0.475

Every additional airport between country with connecting flights from A and B increases the likelihood by 0.167.

As time passes, the likelihood of passing a disease decreases. However, this has as before a small effect.

Table 13: Marginal effects for Probit Regression of African Swine Fever

In the maps of Figure 7 and Figure 8, we look into Brazil and Bolivia as source countries. Even though they're both in the same continent, we can see that effect of neighbouring borders. Brazil is more likely to infect countries in the north of South America that Bolivia cannot due to physical borders.

Both countries have connecting flights to Madrid, Spain and therefore see a similar infection likelihood. When we look at Portugal, Spain's neighbour, we can see that

this is more vulnerable when Brazil is infected. This is due to the fact that Brazil is a previous Portuguese colony and they share several connecting flights.

Figure 7: *[Left]* Probability of infection given that there is an outbreak of yellow fever in Brazil. *[Right]* Probability of infection larger than 0.8 given that there is an outbreak of yellow fever in Brazil.

Figure 8: *[Left]* Probability of infection given that there is an outbreak of yellow fever in Bolivia. *[Right]* Probability of infection larger than 0.8 given that there is an outbreak of yellow fever in Bolivia.

Moving now to Indonesia, we can see that this country is most likely to infect islands around it and Australia as there is a lot of connection among them.

Figure 9: *[Left]* Probability of infection given that there is an outbreak of yellow fever in Indonesia. *[Right]* Probability of infection larger than 0.8 given that there is an outbreak of yellow fever in Indonesia.

When we look into France, we can see this country is more likely to infect others in the European Union, Morocco and Algeria as well as some other countries in Africa and also Brazil. Indeed, France is well connected and economically similar to other Schengen countries which is shown on the map. We also see that historical links of France and its colonial past in Africa (Algeria, Ivory Coast, Congo, etc) shows up on both maps. On the other hand, Indonesia is linked to its respective (non-terrestrial) neighbours : for the non-bordering countries, we expect this to be due to several connecting flights.

Figure 10: *[Left]* Probability of infection given that there is an outbreak of yellow fever in France. *[Right]* Probability of infection larger than 0.8 given that there is an outbreak of yellow fever in France.

6.3.2 Random Forest

At a first glance, Random Forest predicts the best response in terms of accuracy, sensitivity and specificity. The values for these metrics are above 0.9. A summary of these results can be seen in Table 14.

Table 14: Confusion matrix and metrics for Random Forest Classification ('Positive' Class=0)

Yellow Fever

Pred / Actual	0	1
0	24,614	10
1	113	4,137

Accuracy:	0.995
Sensitivity:	0.995
Specificity:	0.997

African Swine Fever

Pred / Actual	0	1
0	11,067	7
1	86	118

Accuracy:	0.991
Sensitivity:	0.992
Specificity:	0.944

However, this is mostly due to the fact that the model heavily overfits the data, offering no room for extrapolation to new data. Indeed, the classification algorithm creates trees so deep that each individual positive outcomes are "remembered" or stored. This has the property of providing apparently stellar results yet we have reasons to believe that those results aren't realistic, the first one being that this is a common issue with RF in the literature. This is particularly worrying as our goal is to use the model on countries that have never been infected by a particular disease. There are some tweaks to make the Random Forest algorithm more robust yet still efficient however we noticed that depending on the dataset, disease and origin countries, we had to adjust the tweaking which makes it very unscalable to a large collection of disease (>140). This is why we decided that although offering a compelling approach from an accuracy standpoint, classification via RF lacked scalability, flexibility and probabilistic interpretation. Another issue *only* comes up when visualising the predictions:

Figure 11: *[Left]* Probability of infection given that there is an outbreak of yellow fever in Brazil. *[Right]* Probability of infection larger than 0.5 given that there is an outbreak of yellow fever in Brazil.

Figure 12: *[Left]* Probability of infection given that there is an outbreak of yellow fever in Bolivia. *[Right]* Probability of infection larger than 0.5 given that there is an outbreak of yellow fever in Bolivia.

First, we see that for Brazil and Bolivia as origin countries, both countries being neighbours and having similar characteristics, we see that the list of potential next targets are alike which is reassuring. However, when picking France and Indonesia, we see that although the countries do not share a common past and have somewhat different characteristics and yet, they respective targets are strikingly identical.

Figure 13: *[Left]* Probability of infection given that there is an outbreak of yellow fever in Indonesia. *[Right]* Probability of infection larger than 0.5 given that there is an outbreak of yellow fever in Indonesia.

However, we observe that none of Indonesia nor France have had a yellow fever alert hence are not part of the training / testing set. Therefore, we visualise a model that overfits the data giving us no real insights.

Figure 14: *[Left]* Probability of infection given that there is an outbreak of yellow fever in France. *[Right]* Probability of infection larger than 0.5 given that there is an outbreak of yellow fever in France.

6.3.3 LogitBoost

LogitBoost is a boosting classification algorithm that performs additive logistic regression minimising the logistic loss (Anon n.d.)

As with Random Forest, we got similar performance (see Table 15), but same restrictions were discovered in the visualisations maps.

This model was again overfitting the data and was predicting commonly infected countries as the most vulnerable.

We hence decided not to pursue with this model.

Table 15: Confusion matrix and metrics LogitBoost ('Positive' Class=0)

Yellow Fever

Pred / Actual

	0	**1**
0	24,655	2
1	662	3,556

Accuracy: 0.977
Sensitivity: 0.974
Specificity: 0.999

African Swine Fever

Pred / Actual

	0	**1**
0	11,081	3
1	123	72

Accuracy: 0.989
Sensitivity: 0.989
Specificity: 0.960

7. Analysis and Results

As mentioned in the previous section, we created a dataset for visualisation purposes. This dataset included every country as source country A, paired to each other country. For each pair, we added their corresponding country metrics. We used 30 days as time difference and used January as the month to calculate the temperatures and rainfall of each country.

With this dataset, we then created a network, in which the nodes were represented by countries. Countries could be connected by an edge if the corresponding likelihood was larger than 0.8. It is important to notice that the resulting network is a directed graph given that the likelihood is calculated based on a specific source country A.

Out-degree of a country A is the sum of likelihoods outcoming links with values larger than 0.8 to countries B. We will refer it as metric to identify countries that are more likely to infect others.

In-degree of country B is sum of likelihoods incoming links with values larger than 0.8 from source countries A. We will use this metric to identify the countries that are more vulnerable to be infected by a different country.

Table 16 shows the top 10 countries with highest levels of in-degree and out-degree. From the first list, we can notice that infectious countries are more likely to be countries that are well connected due to tourism or economic purposes. These are all considered developed countries. It is surprising to see the islands Trinidad and Tobago and Antigua and Barbuda. This is possibly explained by economic strength and tourism. Trinidad and Tobago has the third highest GDP per capita based on purchasing power parity (PPP) in the Americas, strongly influenced by the petroleum industry while Antigua and Barbuda is well known for its luxury resorts attracting many people around the world. Nevertheless, if we look into the out-degree distribution in Table 17, we can notice that most uncommon out-degree scores were the ones above 115.

On the other hand, the top 10 in-degree table shows that the countries most likely to become affected tend to be developing countries with more unstable governments.

When we looked into the top 20 list, we noticed that many countries from the European Union were also included. This may suggest that free borders have a strong impact on the spread of communicable diseases.

Table 16: Top 10 countries with highest levels of “in-degree” and “out-degree” (Yellow Fever)

Country	Out-degree	Country	In-degree
Qatar	184	Madagascar	202
Macao	179	South Sudan	167
United Arab Emirates	139	Mozambique	156
Hong Kong	114	Congo	152
Brunei	109	Haiti	152
Saudi Arabia	107	Eritrea	144
Singapore	89	Marshall Islands	141
Trinidad and Tobago	76	Angola	140
Antigua and Barbuda	74	Gabon	139
Kuwait	71	Nigeria	108

Table 17: *[Left]* Out-degree distribution and *[Right]* In-degree distribution

The images in Figure 15 show the source countries “A” in the y axis and the “B” countries in the x axis.

When we see dark horizontal lines, we are looking into the countries that are more likely to infect many others. For instance, the first top left graph suggests United Arab Emirates and Brunei as very infectious while the bottom left graph suggests Hong Kong.

The vertical lines shows the countries that are more vulnerable to infected by others. The top right graph shows Congo while the bottom right identifies Haiti as more vulnerable.

Figure 15: Yellow Fever likelihood matrix for groups of 30 countries

In the same way, we can look into regions and try to understand the dynamics of infectious diseases.

Figure 16 shows on the left the Eastern European countries as countries A while the on the right we can see the Western European countries as countries A. We immediately notice that Western European countries link to more countries around the world. It is interesting to see that Luxembourg as hub. This may be due to the fact that it is wealthiest country in the European Union, per capita (Anon n.d.).

Figure 16: *[Left]* Filtering Eastern European countries as source countries A. *[Right]* Filtering Western European countries as source countries A.

If we then look into Central and South-Eastern Asia, we can notice a dramatic contrast.

Central Asian countries as Kazakhstan, Uzbekistan and Turkmenistan are more isolated and present lower population densities. We therefore see very few connections. However, if we look into South-Eastern Asia, we see a well connected graph, with clear big participants as Singapore, Thailand and Brunei.

Figure 17: *[Left]* Filtering Central Asian countries as source countries A. *[Right]* Filtering South-Eastern Asian countries as source countries A.

8. Conclusion and recommendations

The aim of this project was to get insight into ProdMED data and explore the possible values that could be created. Whilst the ProdMED dataset we had to work with was small, by augmenting and combining the ProdMED dataset with externality available data and employing supervised machine learning modelling method, we've been able to model conditional probability of infection with Probit Regression to identify most contagious and/or vulnerable countries for an infection.

After trying different models, we found that probit regression model was the most appropriate model used to calculate the probability and simulated disease transmission. From the model we could see that some countries were more likely to infect others while other countries were more vulnerable to be infected by others.

For yellow fever, the countries that were more likely to infect others were usually wealthy and well connected by air traffic. Countries that were more vulnerable to contract an infectious disease usually included developing nations with lower levels of public health and socio-economic and political stability. Nevertheless, we could see that Schengen countries were more vulnerable compared to other wealthy countries, suggesting that open borders may have a strong impact on the spread of communicable diseases.

We recommend ProMED to use these findings to identify most infectious countries and most vulnerable countries to be infected. In this way, they can inform their partners on how to best focus their resources on such countries in order to prevent the spread of these diseases.

Using the probit model, ProMED can also make prediction on possible changes as the improvement of public health or political stability. This can provide data evidence to take action in certain areas that can be developed.

References

ProMed, (2010) ProMed Official Website [online] Available at: https://www.promedmail.org/aboutus/ (Accessed by 30 March, 2018)

Anon, LogitBoost Classifier. *Developer Guide for Intel® Data Analytics Acceleration Library 2017 Update 3*. Available at: https://software.intel.com/sites/products/documentation/doclib/daal/daal-user-and-reference-guides/daal_prog_guide/GUID-E9524C96-C514-45A2-838B-F10DE6446FE0.htm [Accessed April 1, 2018a].

Anon, Luxembourg. *U.S. News*. Available at: https://www.usnews.com/news/best-countries/luxembourg [Accessed April 1, 2018b].

Anon, SpaCy. *SpaCy - Industrial-Strength Natural Language Processing*. Available at: https://spacy.io/ [Accessed March 28, 2018c].

Eisenstein, M., 2018. Infection forecasts powered by big data. *Nature*, 555(7695), pp.S2–S4.

Held, L., Meyer, S. & Bracher, J., 2017. Probabilistic forecasting in infectious disease epidemiology: The thirteenth Armitage lecture. Available at: http://dx.doi.org/10.1101/104000.

Kenyon, T.A. et al., 1996. Transmission of multidrug-resistant Mycobacterium tuberculosis during a long airplane flight. *The New England journal of medicine*, 334(15), pp.933–938.

Metcalf, C.J.E. et al., 2009. Seasonality and comparative dynamics of six childhood infections in pre-vaccination Copenhagen. *Proceedings. Biological sciences / The Royal Society*, 276(1676), pp.4111–4118.

Moore, M. et al., 2017. Identifying Future Disease Hot Spots: Infectious Disease Vulnerability Index. *Rand health quarterly*, 6(3), p.5.

Mossong, J. et al., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3), p.e74.

Shaman, J., 2016. Forecasting Infectious Disease Outbreaks.

Woolhouse, M., 2011. How to make predictions about future infectious disease risks. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 366(1573), pp.2045–2054.

Tatem, A. J., Rogers, D. J., & Hay, S. I. (2006). Global transport networks and infectious disease spread. *Advances in parasitology*, *62*, 293-343.

WHO, 2016. WHO reports [online] http://www.who.int/en/ [Accessed by 28 March,

2018]

Reich, N. et.al., 2017. Quantifying the Risk and Cost of Active Monitoring for Infectious Diseases. *Scientific Reports,* 19, 12-13

.